Open Source XML Database Toolkit:

Resources and Techniques for Improved Development

Liam Quin

Wiley Computer Publishing

John Wiley & Sons, Inc.

NEW YORK · CHICHESTER · WEINHEIM · BRISBANE · SINGAPORE · TORONTO

Publisher: Robert Ipsen

Editor: Cary Sullivan

Assistant Editor: Christina Berry

Managing Editor: Marnie Wielage

Associate New Media Editor: Brian Snapp

Text Design & Composition: Pronto Design, Inc.

This book is printed on acid-free paper. ♾

Published by John Wiley & Sons, Inc.

Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:
Quin, Liam.
Open source XML database toolkit : resources and technniques for improved development / Liam Quin.
p. cm.
ISBN 0-471-37522-5 (pbk. : alk. paper)
1. XML (Document markup language) 2. Web databases. I. Title.

QA76.76.H94 Q56 2000
005.7'2—dc21 00-033022
CIP

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

ACKNOWLEDGMENTS

This book has been over a year in the making. During that time, the world changed: XML became popular; Perl and Python got XML support; Oracle announced support for Linux; books and magazines on XML, Linux, and Open Source software sprouted like teenagers. Originally, I had thought I would spend most of the book describing tools, but when the tools kept changing, I abandoned that plan, and concentrated on techniques.

Through all this, my editor at Wiley, Christina Berry, remained calm and helpful. I changed jobs; my partner Clyde went back to university; and work on the book dragged on. So thanks are due to Christina, and to Clyde, who insisted that I write this book in the first place. I also need to thank Christopher Cashell for his help with the "Resource Guide", Jerji (Jerry Herbert) for distracting me when I was close to despair, and, above all, Moonkitty and Cosmos for purring even in the darkest nights.

INTRODUCTION

Welcome to *The Open Source XML Database Toolkit*. This book is not just for people who write code, but also for those who design and specify programs. There are chapters to introduce XML and databases of various types, and chapters to explain how to use them together, all in an open source environment.

Open Source

This book, as is obvious from its title, is about open source software. Most of the tools discussed are open source. Open source software is software that is distributed with its program source code, under a license that allows you to modify and redistribute that source code. In other words, if you don't like the way the program works, *you are free to change it*. People sometimes think that open source software refers to software that is free of charge, but that's not the case. Software in the context of open source is software that you are free to change and to use in any way you wish, as long as you don't try to restrict that freedom from other people.

That said, note that most open source software *is* distributed more or less free of charge, with development costs paid for by support fees, services, or even donations (there may be media and shipping charges, of course).

Also note that this book does mention software that is *not* open source, usually when the software is very significant or when there are no open source competitors. Even if the most appropriate tool is not free, it's better to use it than to be held back by ideology or dogma. Here are some examples of nonfree software mentioned in this book:

Oracle (www.oracle.com). The most widely used high-end commercial relational database.

Solaris (solaris.sun.com). The Sun Solaris operating system running on a Sun SPARC server is probably the best-engineered and most stable server platform today. The source code to Solaris *is* available for a small fee, but the license is restrictive.

SoftQuad XMetal (www.softquad.com). XMetal is a widely used editor for XML documents. It offers a documentlike interface, a structured interface, and a source code view. However, it's only available to run under Microsoft Windows.

Object Design's ObjectStore (www.excelon.com; www.odi.com). ObjectStore is one of the better-known object-oriented databases. Although there is a free version for Java (PSE), source is not available, and there are restrictions on its use.

Most of the software described in this book is free. In a few cases, you may need to pay royalties if you use the software as part of a product or service that you sell, so be sure to check the licenses. Here are a few examples of free software mentioned in this book:

FreeBSD (www.freebsd.org). FreeBSD is one of several open source and free operating systems mentioned in this book. The examples were tested on FreeBSD and Linux.

MySQL (www.mysql.com). MySQL is a freely available relational database, although royalties may apply for some uses. It lacks many of the features of a high-end commercial system such as Oracle's, but it is very widely used and fast, and will take you a long way.

XT (www.jclark.com). XT is an implementation of the XML Style Language Transformation specification, which is a complicated way of saying that it manipulates XML documents, for example to produce HTML or XHTML.

Apache (www.apache.org). Apache is the most widely used Web server on the Internet. In addition to being open source and free, it is also very powerful and robust.

NOTE

Go to www.opensource.org for more information about the open source movement.

XML

The eXtensible Markup Language, XML, is a way of defining simple text-based representations of arbitrarily complex structured information. The term XML is also used to refer to data that's marked up in a format defined using XML.

This is a book about working with XML. You might have XML documents that you need to store in a database, or you might want to use XML as an interchange format.

Chapter 1 in Part One, "Just Enough XML", introduces the main concepts of XML.

Database

This is also book about using databases. You'll find an introduction to the Structured Query Language, SQL, in Chapter 3, "Just Enough SQL."

The book doesn't only address relational databases, though. You'll find descriptions of hashing (Chapter 12: "Dynamic Hashing: *ndbm*"), of object-oriented databases (Chapter 10, "Introduction to Object-Oriented Databases," and Chapter 11, "XML as Classes and Objects"), and of text retrieval databases (Chapter 13, "Text Retrieval Technology Overview").

In all cases, the emphasis is on using XML and databases together in an open source environment.

Toolkit

As in all the best toolkits, there are lots of toys to play with. Most of them are listed for reference in Part Five, "Resource Guide." The toolkit approach means that this book does not go deeply into any single product or tool, but instead focuses on using tools together. Chapter 2, "Client/Server Architecture," introduces network programming, but from then on, the idea of using applications together pervades the book.

The power lies in the way the tools work together. Not only are these open source tools, meaning you can change them to make them work together in the way you want, but they are also widely used and powerful tools, meaning you probably won't have to change them.

Welcome to the open source revolution.

About the Illustrations

The illustrations in printed books generally have to use crisp lines, as if everything was polished and perfect. But don't be deceived by this. A quick sketch on the back of an envelope, or on a whiteboard, can help people to understand the relationships between components where a textual description cannot. Never hesitate to draw pictures, and don't worry if they are not very polished.

Typographic Conventions

I've tried to keep things simple in this regard. In the few places where I've shown a session at an interactive terminal or shell, the prompt is given as the pound sign () if you need to be logged in as root, and as the dollar sign ($) otherwise. The text you type is in bold. Here's a brief example:

```
$ pwd
/export/home/liam
$
```

The dollar sign on the third line shows the prompt after the command (pwd, print working directory) completed.

If you are thinking that this looks suspiciously like a Unix (or Linux) shell, you're right. If you're using Microsoft Windows, don't despair—there's a lot you can learn from this book. But if you want to write reliable high-performance database applications, you should develop them on Unix if you can. Imagine going for a whole year of development without a single machine crash and you'll see why.

Source code is shown with function names in **bold** when they are defined. This is just for your convenience, since books don't have a search command. If you type the examples into the computer, or download them, they are plain text, with no formatting. In the same way, comments are shown in *italics.*

What's on the Web Site?

At the companion Web site for this book (www.wiley.com/compbooks/quin) you can find:

- The complete text of the "Resource Guide" in HTML, with links to all of the resources mentioned.
- All of the examples and source code from the book, along with complete or enhanced examples.
- The data for the BookWeb example, along with a simple shell script to create the sample database using MySQL under Linux/Unix.
- The BookWeb site.
- AutoLinker, with the Glossary and the Dictionary examples.
- The sample Web server, written in Perl.

Other XML resources are added from time to time, and the "Resource Guide" is updated occasionally. And note, you may need your copy of the printed book ready before downloading the examples.

Finally, feel free to contact me (liam@holoweb.net; http://www.holoweb.net/~liam/); I'm always interested in comments and suggestions for the next edition. You can also find me on the SorceryNet Internet Relay Chat network (irc.sorcery.net; its Web site is at www.sorcery.net).